Spectral Regularization in Generalized Matrix Learning Vector Quantization

David Nova
Millennium Institute of Astrophysics &
Department of Electrical Engineering, University of Chile, Chile
Email: dnovai@ug.uchile.cl

Pablo A. Estévez
Department of Electrical Engineering, University of Chile &
Millennium Institute of Astrophysics, Chile
Email: pestevez@ing.uchile.cl

***Abstract*—In this contribution we propose a new regularization method for the Generalized Matrix Learning Vector Quantization classifier. In particular we use a nuclear norm in order to prevent oversimplifying/over-fitting and oscillatory behaviour of the small eigenvalues of the positive semi-definite relevance matrix. The proposed method is compared with two other regularization methods in two artificial data sets and a real-life problem. The results show that the proposed regularization method enhances the generalization ability of GMLVQ. This is reflected in a lower classification error and a better interpretability of the relevance matrix.**

I. Introduction

Learning vector quantization (LVQ) is a prototype-based classifier belonging to the family of k-nn classifiers [1]–[3]. The prototypes aim at representing class regions of the input space formed by intersection of hyperplanes and producing a Voronoi tessellation. LVQ has been extensively used in part due to its easier interpretability which is an advantage over black box classifiers such as multilayer perceptrons (MLP) and support vector machines (SVM).

The main drawback of LVQ is that it is metric dependent. In machine learning, an appropriate way to measure the distance is needed. This has resulted in a new interesting topic in machine learning the so-called metric learning [4]. Metric learning aims at automatically learning a metric from data. The key properties of metric learning algorithms are: learning paradigm, form of metric, scalability, optimality of the solution and dimensionality reduction [4]. In [5] a metric learning model was proposed in order to deal with this issue known as Generalized Matrix Learning Vector Quantization (GMLVQ). It uses a positive semi-definite matrix to project the data into a manifold and extracts the relevance of features from data in a supervised way, allowing a limited rank representation of the relevance matrix [6].

Basically, a metric learning yields the matrix which best fits the underlying semantic of the data, i.e., the relevant information of the data that helps enhance the classification performance. This is typically formulated as an optimization problem with a regularization term. The latter reduces the effect of oversimplifying or over-fitting depending on the value of the regularization parameter. Oversimplifying occurs when the solution is too sparse and some relevant information to the classification task is lost. On the contrary, over-fitting means that the solution is extremely adjusted to the training data, which results in a poor generalization ability of the classifier. A regularization study for GMLVQ was carried out in [7], where a regularization term was proposed. The authors found that GMLVQ displays a tendency towards oversimplification during training, instead of over-fitting, contrary to conventional classifiers. In particular, it tends to eliminate too many dimensions (super sparse solution), affecting negatively the classification performance. In [8] a study of different penalty functions was performed in order to analyse the behaviour of the metric learning algorithm and the resulting matrix relevance. Both studies aimed at enhancing the generalization ability of the GMLVQ model.

In the last years many researchers in machine learning have used a spectral regularization [9] in order to deal with noisy data and prevent over-fitting. This technique has many applications for example trending in TV programs [10] (the "Netflix" competition). Originally, spectral regularization was proposed in order to deal with ill-posed inverse problems. In particular spectral regularization helps prevent the oscillatory behaviour associated with small eigenvalues. In addition it gives a low rank approximation of the matrix.

In this contribution we propose a spectral regularization of the GMLVQ algorithm by using a nuclear norm to deal with over-simplifying in the metric learning process. Besides, we describe the properties of the nuclear norm and its highlights. We validate our proposal by comparing it with two different regularization methods for GMLVQ. The first one was proposed in [7] where the determinant of the relevance matrix was considered. The second one is the Smoothly Clipped Absolute Deviation (SCAD) [11] which application to GMLVQ was previously studied in [8]. All comparisons were done in two artificial and a real-life data set. As metric performance we use classification rate and the interpretability of the relevance matrix.

The remainder of this paper is organized as follows: In section 2, background concepts are introduced. In section 3, the proposed method is described. In section 4, the experiments are described and the results for the three data sets are shown. Finally, in section 5, the conclusions are drawn.

II. Background

In this section some relevant concepts are described before introducing the proposed method.

978-1-5090-6638-4/17/$31.00 ©2017 IEEE

A. Metric Learning

Metric Learning is a branch of machine learning that aims to learn automatically a metric from data. In classification tasks the choice of the metric plays an important role because it is reflected in the generalization ability and classification error. Thus, the goal of metric learning is to adjust a real-valued metric function.

In [4] the metric learning properties are described. The learning paradigm could be fully supervised, weakly supervised or semi-supervised. The form of metric can be linear, non-linear or local. Scalability can be with respect to the number of samples or dimensions. Optimality of the solution can be local or global. Dimensionality reduction can be used or not. Let X be a data matrix. For a weakly supervised way the problem is formulated by using a positive semi-definite (PSD) matrix as follows:

$$\min_{\mathcal{L}} \ell(\mathcal{L}, \mathcal{S}, \mathcal{D}, \mathcal{R}) + \beta R(\mathcal{L}), \tag{1}$$

where $\ell(\cdot)$ is a loss function, $\mathcal{L}$ is a PSD matrix, $\mathcal{S}$ is a set of must-link pairs (pairs of x_i and x_j should be similar), $\mathcal{D}$ is a set of cannot-link pairs (pairs of x_i and x_j should be dissimilar), $\mathcal{R}$ is a set of relative constraints, $R(\cdot)$ is a regularization function and β is the regularization parameter. A learned metric is generally used in k-nn classifiers or some clustering methods such as K-means. The metric learning can be embedded in the classifier as in the case of GMLVQ.

B. Generalized Matrix Learning Vector Quantization

GMLVQ introduces a matrix relevance in order to learn a metric during the training process. It corresponds to an embedded metric learning into a prototype-based classifier. Let $X \in \mathbb{R}^{N\times n}$ be a data matrix where N is the number of samples and n the number of dimensions, $\Lambda \in \mathbb{R}^{n\times n}$ the relevance matrix and $W \in \mathbb{R}^{M\times n}$ a matrix of prototypes, where M indicates the number of prototypes and n the dimensionality. The GMVLQ cost function is as follows:

$$E = \frac{1}{N}\sum_{i=1}^{N} f\left(\mu\left(x_i, W, \Lambda\right)\right), \tag{2}$$

where $\mu(\cdot) \in [-1, 1]$ is the relative distance function

$$\mu(x) = \frac{d^+ - d^-}{d^+ + d^-}, \tag{3}$$

where d^+ is the distance of the closest correct class prototype to x, and d^- is the distance of the closest wrong class prototype to x. Also, the following sigmoid function is defined

$$f_\tau(a) = \frac{1}{1 + e^{-a\cdot\tau}}, \tag{4}$$

where parameter τ is usually increased with the number of training epochs during the learning process. When τ is increased $f_\tau(a)$ the cost function in (2) approximates the classification error. Furthermore, with a large value of τ the prototypes become sensitive to the border of the class distributions. Here Λ is a PSD matrix and can be decomposed as $\Lambda = \Omega^T\Omega$ with $\Omega \in \mathbb{R}^{n\times k}$, where $k \le n$ is the intrinsic dimension of the manifold, which satisfies:

$$z^T\Lambda z \ge z^T\Omega^T\Omega z \ge 0, \tag{5}$$

i.e., all eigenvalues of Λ are positive. The metric or similarity measure is defined as:

$$\begin{aligned} d_\Lambda(x, y) &= (x-y)^T\Lambda(x-y) \\ &= (x-y)^T\Omega^T\Omega(x-y). \end{aligned} \tag{6}$$

This metric implicitly corresponds to computing the square Euclidean distance after a linear projection of the data defined by the transformation Ω. Being Λ a real PSD matrix it guarantees the properties of the pseudo-distance: non-negativity, identity, symmetry and triangle inequality as well. Note that if Λ is a low-rank matrix, i.e., rank$(\Lambda) = k < n$, then it yields a linear projection of the data into a space of lower dimension k.

C. Spectral Regularization

Regularization is associated with introducing additional information in order to solve an ill-posed problem or to prevent over-fitting of the model. Spectral regularization [9] works on the eigenvalues of the PSD matrix $\mathcal{L}$. Using a matrix norm which considers the eigenvalues information of the PSD matrix, such as the nuclear norm (or trace norm), we can obtain a regularization over the eigenvalues of the PSD matrix. The nuclear norm belongs to the family of the Schatten norms and it is defined as follows:

$$||A||_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(A)\right)^{1/p}, \tag{7}$$

where $\sigma(A)$ denotes the vector of decreasingly ordered singular values and $\sigma_i(A)$ denotes the i^{th} largest singular value of A, and $\{m, n\}$ corresponds to the number of rows and columns of A, respectively. This norm is sub-multiplicative, i.e. $||AB|| \le ||A||||B|| \;\forall\; A$ and $B \in \mathbb{R}^{n\times n}$. In addition, it is unitarily invariant, i.e. $||A|| = ||UAV||$ for all A and all unitary matrices U and V [12]. When $p = 1$ the nuclear norm results in:

$$\begin{aligned} ||A||_* &= \text{trace}\left(\sqrt{A^*A}\right) \\ &= \sum_{i=1}^{\min\{m,n\}} \sigma_i(A), \end{aligned} \tag{8}$$

where A^* denotes the conjugate transpose, and $\sqrt{A^*A}$ is well-defined and it is another PSD matrix. Note that the nuclear norm is the sum of the singular values of A, i.e., the roots of the eigenvalues of A^*A.

III. Spectral Regularization in GMLVQ

In order to prevent oversimplifying/over-fitting in GMLVQ we propose a nuclear norm regularization of (2) as follows:

$$E = \frac{1}{N}\sum_{i=1}^{N} f\left(\mu\left(x_i, W, \Lambda\right)\right) + \beta||\Omega||_*. \tag{9}$$

Note that depending on the value of β we can deal with overfitting ($\beta > 0$) or oversimplifying ($\beta < 0$). The nuclear norm is not differentiable, however, a smooth approximation was proposed in [13]. We consider convex and non-convex smooth approximations to the rank function. Let's define the following smooth Schatten-p function:

$$\begin{aligned} f_p(A) &= \text{trace}(A^TA + \gamma I_k)^{p/2} \\ &= \sum_{i=1}^{n}(\sigma_i^2(A) + \gamma)^{p/2}, \end{aligned} \tag{10}$$

where $\mathbf{I}_k$ is the identity matrix of size $k \times k$. Note that f_p is differentiable for $p > 0$ and f_p is convex for $p \geq 1$. This approximation converges faster than minimizing directly the nuclear norm function as was demonstrated in [13]. With $\gamma = 0$ and $p = 1$ we have

$$f_1(A) = \|A\|_* \, .$$

Then, the smooth nuclear norm is defined as

$$\begin{aligned} f_1(A) &= \text{trace}(A^TA + 0 \cdot I)^{1/2} \\ &= \text{trace}(A^TA)^{1/2} \end{aligned} \tag{11}$$

and the gradient of the smooth nuclear norm [14], [15], is the following:

$$\begin{aligned} \nabla f_1(A) &= 1 \cdot A(A^TA + 0 \cdot I)^{(1/2)-1} \\ &= A(A^TA)^{-1/2} \\ &\approx A((A^TA)^{1/2})^+, \end{aligned} \tag{12}$$

where $((A^TA)^{1/2})^+$ denotes the pseudo-inverse matrix of $(A^TA)^{1/2}$. Note that the square root of A^TA is another PSD matrix which implies that it is not invertible if there were null eigenvalues. Also, when Ω is a squared matrix, the singular values are the square root of the eigenvalues.

A. Update rules

The stochastic gradient descent is used for minimizing the cost function in (9). For each sample presented to the GMLVQ two prototypes are updated: the winner prototype of the right class and the winner prototype of the wrong class, as follows:

$$w^j = w^j - \alpha\frac{\partial E}{\partial w_j}, \; j = +, - \tag{13}$$

where $+$ indicates the nearest (winner) prototype of the correct class and $-$ indicates the nearest (winner) prototype of the wrong class.

Remember that in GMLVQ the distance is modified as shown in (6). Therefore, the distance between a sample x and a prototype w is the following:

$$d_\Lambda(x_i, w) = (x_i - w)^T\Omega^T\Omega(x_i - w). \tag{14}$$

The derivative of d_Λ with respect to w yields

$$\begin{aligned} \nabla_w d_\Lambda &= -2\Lambda(x - w) \\ &= -2\Omega^T\Omega(x - w). \end{aligned} \tag{15}$$

Thus, we get the update equations:

$$w^+ \leftarrow w^+ - \alpha \cdot 2 \cdot \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial d^+}\Omega^T\Omega(x - w^+) \tag{16}$$

$$w^- \leftarrow w^- + \alpha \cdot 2 \cdot \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial d^-}\Omega^T\Omega(x - w^-). \tag{17}$$

The $\frac{\partial f}{\partial \mu}$ is the partial derivative of the sigmoid function and works as an adaptive learning rate. Also, $\frac{\partial \mu}{\partial d^+} = \frac{2d^-}{(d^+ - d^-)^2}$ and $\frac{\partial \mu}{\partial d^-} = \frac{-2d^+}{(d^+ - d^-)^2}$. Note that the update rules in Eqs. (16)-(17) push the closest correct prototype towards the current data point x and the closest wrong prototype away from the current data point x.

In the case of the matrix relevance by using stochastic gradient descent the following expression is obtained:

$$\nabla E_{\Omega_{lm}} = \frac{\partial E}{\partial \Omega_{lm}}\frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial d^\Lambda}\frac{\partial d^\Lambda}{\partial \Omega_{lm}}, \tag{18}$$

where the derivative of d^Λ with respect to a single element of Ω_{lm} gives

$$\begin{aligned} \frac{\partial d^\Lambda}{\partial \Omega_{lm}} &= \sum_j (x_m - w_m)\Omega_{lj}(x_j - w_j) + \\ &\quad \sum_i (x_i - w_i)\Omega_{li}(x_m - w_m) \\ &= 2 \cdot (x_m - w_m)\left[\Omega(x - w)\right]_l, \end{aligned} \tag{19}$$

where indices (l, m) denote components of the Ω matrix. Then, the update rule for the matrix element Ω_{lm} is:

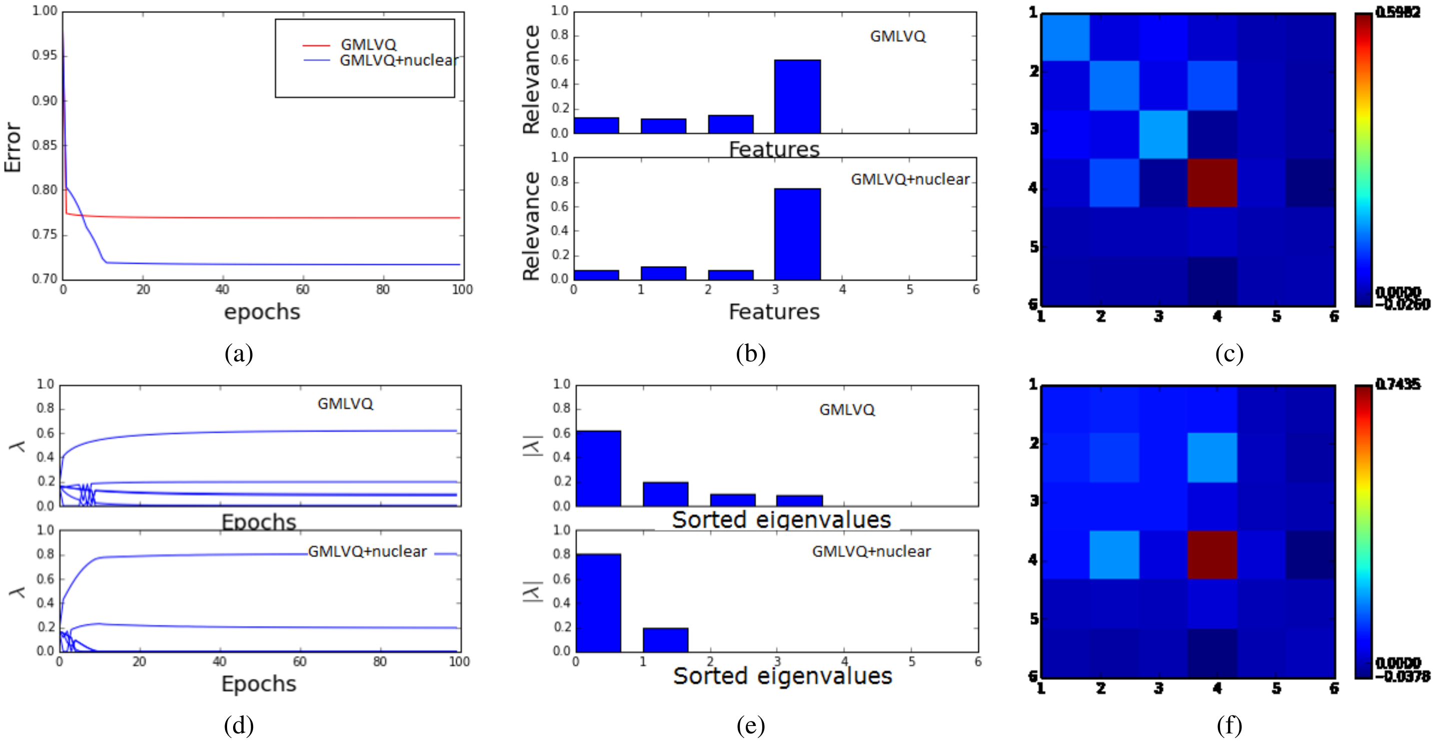
$$\begin{aligned} \Omega_{lm} \leftarrow \; & \Omega_{lm} - \epsilon \cdot 2 \cdot \frac{\partial f}{\partial \mu} \cdot \\ & [\frac{\partial \mu}{\partial d^+} \cdot \left(\left(x_m - w_m^+\right)\left[\Omega\left(x - w^+\right)\right]_l\right) - \\ & \frac{\partial \mu}{\partial d^-} \cdot \left(\left(x_m - w_m^-\right)\left[\Omega\left(x - w^-\right)\right]_l\right)] \\ & +\beta\frac{\partial ||\Omega||_*}{\partial \Omega_{lm}}. \end{aligned} \tag{20}$$

In this way, the learning metric is embedded into the classifier, allowing the relevance matrix to change in such a way that the distance from the closest correct prototype is decreased, whereas the distance from the closest wrong prototype is increased.

Finally, using the smooth approximation of the nuclear norm, defined in (12), and calculating the gradient of this approximation, as defined in (11), the following update rule for the Ω matrix is obtained

$$\begin{aligned} \Omega_{lm} \leftarrow \; & \Omega_{lm} - \epsilon \cdot 2 \cdot \frac{\partial f}{\partial \mu} \cdot \\ & [\frac{\partial \mu}{\partial d^+} \cdot \left(\left(x_m - w_m^+\right)\left[\Omega\left(x - w^+\right)\right]_l\right) - \\ & \frac{\partial \mu}{\partial d^-} \cdot \left(\left(x_m - w_m^-\right)\left[\Omega\left(x - w^-\right)\right]_l\right)] + \\ & \beta\Omega((\Omega^T\Omega)^+)^{1/2}\frac{\partial \Omega}{\partial \Omega_{lm}}, \end{aligned} \tag{21}$$

Fig. 1. Visualization of the results for Toy data set for the original GMLVQ and GMLVQ with nuclear norm regularization (with $\beta = 0.5$). The evolution of the classification error cost function during training (a). The relevance profile after training (b) for GMLVQ (top) and GMLVQ+nuclear (bottom). Visualization of the relevance matrix Λ for GMLVQ (c). Evolution of the eigenvalues (d) for GMLVQ (top) and GMLVQ+nuclear (bottom). The final eigenvalue profile (e) for GMLVQ (top) and GMLVQ+nuclear (bottom). Visualization of the relevance matrix Λ for GMLVQ+nuclear (f).



(a)

(b)

(c)

(d)

(e)

(f)

where $\frac{\partial \mathbf{\Omega}}{\partial \mathbf{\Omega}_{lm}} = \mathbf{J}^{lm}$, and $\mathbf{J}^{lm}$ is a single-entry matrix: 1 at entry (l, m) and zero elsewhere. Also, the matrix pseudo-inverse was used in order to prevent undefined operations when $\Lambda = \Omega^T \Omega$ is not invertible. After each update Λ_{ii} should be normalized to prevent the algorithm from degeneration such as follows:

$$\sum_i \Lambda_{ii} = 1. \tag{22}$$

The eigendirections of Λ are the temporary coordinate system with relevances Λ_{ii} such as in [5]. Note that

$$\sum_i \Lambda_{ii} = \sum_k \Omega_{ik} \Omega_{ki} = \sum_k (\Omega_{ik})^2. \tag{23}$$

Therefore, a normalization takes place by dividing all elements of Ω with the following expression:

$$\left(\sum_{ik} (\Omega_{ik})^2 \right)^{1/2} = \left(\sum_i [\Omega \Omega]_{ii} \right)^{1/2}. \tag{24}$$

Otherwise, the Ω matrix might increase its values and the metric learning might diverge.

IV. Experiments

In order to validate our regularization approach we have compared it with other regularization methods. The first one SCAD [11] considers a regularization by using an approximation of the l_0 function and regularizing over each element of the Ω matrix, i.e., it does not work over full matrix operations. The second one [7] , log regularization, considers the determinant of the relevance matrix, working over a full matrix. For all experiments a 10-fold cross validation was done. Besides, an ANOVA test with Bonferroni correction ($\alpha = 0.05$) was performed in order to determine statistical differences among the results in terms of the classification performance.

In what follows we denote the main diagonal of the relevance matrix Λ as the relevance profile, and the eigenvalues of Λ as the eigenvalue profile. In addition, we refer to GMVLQ with SCAD regularization as GMLVQ+SCAD, GMLVQ with log regularization as GMLVQ+log, and GMLVQ with nuclear norm regularization as GMLVQ+nuclear.

Note that the classification error given in Eq. 2 is in the range $[0, N]$ due to the use of the sigmoid function. The first term on the right side of Eq. 9 is normalized by the number of samples N, so that the classification error takes values in $[0, 1]$. In this way the classification error becomes independent of the number of samples. A perfectly right classifier with a large τ value would yield a nil error. On the other hand, a completely wrong classifier would yield a unitary normalized classification error. The second term on the right hand of Eq. 9 corresponds to the nuclear norm of the matrix Ω multiplied by the regularization parameter β. The nuclear norm is the sum of the singular values of a matrix. Note that in our procedure, we normalize the Ω matrix in such a way that its trace is bounded above to 1. The regularization parameter controls the trade-off between the first and the second term. Its value restricted to $|\beta| < 1$, but usually is a small number so that the classification error is more important. Taking into account the aforementioned we selected the β value by trial and error.

A. *Toy data set*

This data set is a continuous version of the XOR problem (2 classes) with 4 normal distributions. Six features are defined. Feature 4 is relevant, features 1-3 are redundant versions with additive noise, and features 5-6 are irrelevant. In this experiment 2 prototypes per class obtained the best results. Here for didactic purposes we study the behaviour and importance of using regularization and how the proposed regularization method can enhance the GMLVQ classifier without losing relevant information. As mentioned before there are two extreme scenarios: over-fitting and oversimplifying in the matrix relevance Λ.

Fig. 1 shows an example when the regularization parameter was set to $\beta = 0.5$. It can be observed that the proposed regularization method deletes the small eigenvalues contaminated with noise. Fig. 1 (d) shows the evolution of the eigenvalues. It can be observed that the proposed regularization deletes small eigenvalues in the early stage of learning (before 20 epochs) which is shown also in the eigenvalue final profile Fig. 1 (e). Fig. 1 (b) shows a visualization of the main diagonal of Λ as a relevance profile for GMLVQ without (top) and with (bottom) regularization. The regularization relevance profile shows an increase in the values of the most relevant features and a decrease in the values of the redundant features (1-3). Fig. 1 (c) and (f) show the visualization of the Λ matrix obtained with and without regularization, respectively. It can be observed that both visualizations are similar which means that no relevant information has been lost. Fig. 1 (e) shows that the small eigenvalues are eliminated. The eigenvalue profile shows that the problem can be solved in two dimensions because only two eigenvalues are different from zero when using regularization.

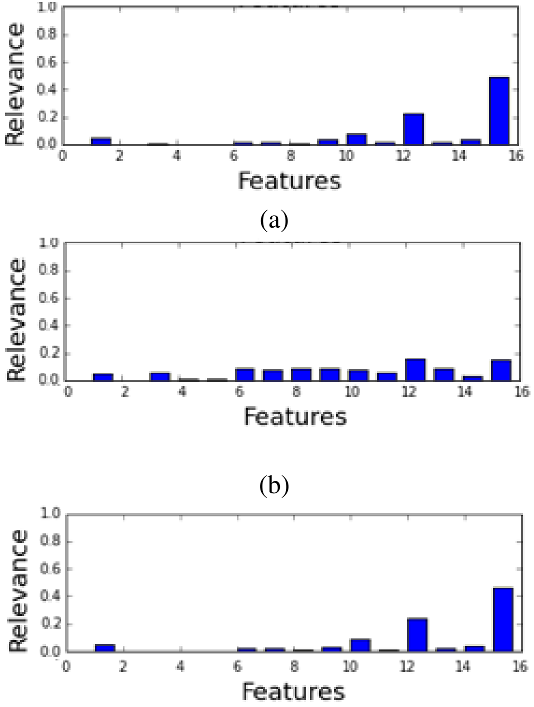
B. *Pipeline data set*

The Pipeline data set consists of 12 dimensions and 3 classes. Here 3 prototypes per class gives the best results for this data set. The results are shown in Table 1. All regularization methods are statistically different with respect to the original GMLVQ without regularization but we cannot distinguish among the regularization methods for the Pipeline data set.

C. *Image Segmentation data set*

The Image Segmentation data set consists of 19 features (3 were eliminated for being constant), 2100 samples and 7 classes. In this experiment 3 prototypes gives the best classification performance.

Fig. 2. Relevance profiles for different β values using the nuclear norm regularization on the Image Segmentation data set. Relevance profiles with $\beta = 0$ (a), $\beta = -0.5$ (b) and $\beta = 0.5$ (c).



(a)

(b)

(c)

Fig. 2 shows the influence of the regularization term. Fig. 2 (a) corresponds to the original GMLVQ classifier relevance profile (main diagonal of the relevance matrix Λ) without using regularization, Fig. 2 (b) corresponds to the relevance profile with $\beta < 0$, and Fig. 2 (c) corresponds to the relevance profile with $\beta > 0$. We can observe that the relevance profile may go from being rather uniform ($\beta < 0$) to very sparse ($\beta > 0$) in comparison with the case $\beta = 0$. When using a $\beta < 0$ value the nuclear norm helps prevent oversimplification of the problem and allows preserving the most relevant eigenvalues for the classification. For $\beta > 0$, the nuclear norm helps prevent over-fitting and deletes noisy eigenvalues as well. But it could delete features that may not have a high relevance value but when interacting with other features they may add complementary information to the classification.

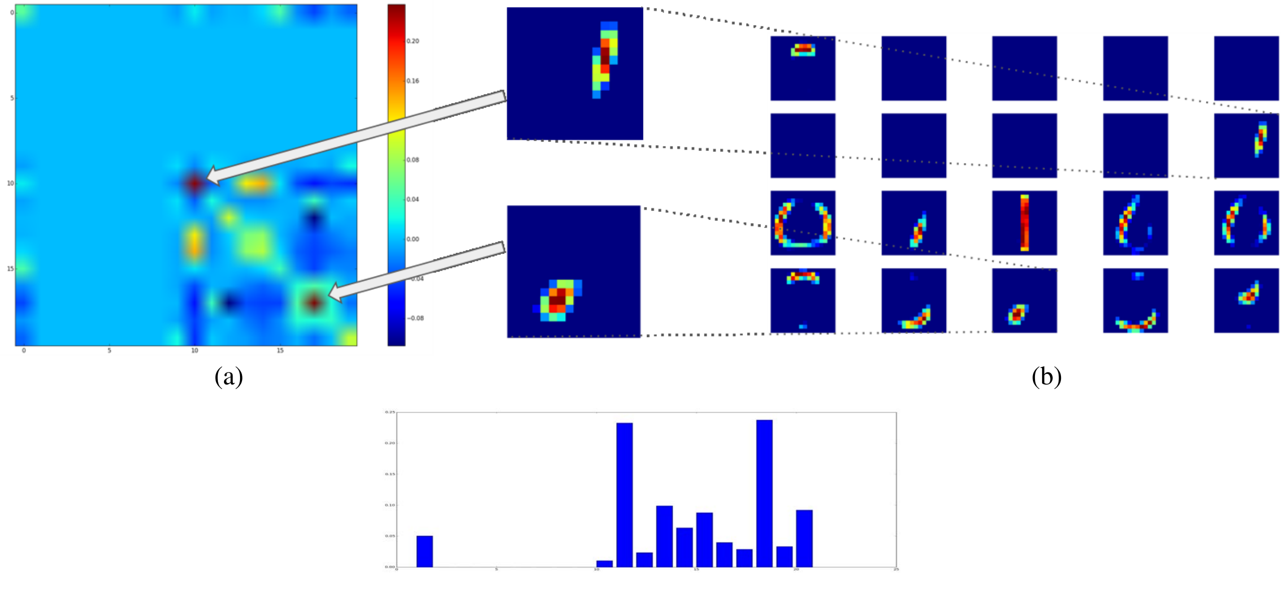
In terms of the classification performance both GMLVQ+log and GMLVQ+nuclear are the best in this experiment (such as is summarized in Table 1). This is because these methods use information about all the relevance matrix in contrast with GMLVQ+SCAD. The latter uses information from just single elements of the data matrix losing information about of eigenvalue profile of the relevance matrix and important information is lost. In this sense, the nuclear norm uses information about the eigenvalue profile and helps preserve elements with low relevance value if they contribute by interacting with other features. Moreover, there is no significant statistical difference between the results obtained with GMLVQ and GMLVQ+SCAD in this experiment.

D. *Real world data set (USPS)*

The United States Postal Service (USPS) image data set consists of handwritten digits from 0 to 9. It has 7291 training observations and 2007 test observations, which are 16x16 gray scale images.

In this data set, we did a preprocessing step by using non-negative matrix factorization (NMF) in order to reduce the dimensionality of the data set [16], [17]. We used different k values for projecting the original data set in manifolds of different dimensions and the best performance was obtained for $k = 20$. This new projection was used as input data for all models compared in this experiment. In this way each element used for training GMLVQ represents a non-negative projection

Fig. 3. A schematic of the matrix relevance and its relationship with the dictionary matrix obtained by Non Negative Matrix Factorization for $k = 20$ elements. The relevance matrix Λ and the highlighted patterns (a). The dictionary resulted after applying NMF for $k = 20$ (b). The relevance profile of the relevance matrix is shown in bar plot (c).



(a)

(b)

(c)

TABLE I. CLASSIFICATION ERROR RATE (MEAN ± STD) FOR THE ORIGINAL GMLVQ AND THREE REGULARIZATION METHODS.

	GMLVQ	GMLVQ+SCAD	GMLVQ+Log	GMLVQ+nuclear
Pipeline	0.0600 ± 0.0085	0.0020 ± 0.0133	0.0022 ± 0.0063	0.0018 ± 0.0031
Image Segmentation	0.0405 ± 0.0125	0.0310 ± 0.0139	0.0233 ± 0.0006	0.0195 ± 0.0008
USPS	0.1321 ± 0.0837	0.1022 ± 0.0025	0.0951 ± 0.0050	0.0901 ± 0.0035
USPS ($k = 20$)	0.0622 ± 0.0331	0.0435 ± 0.0220	0.0350 ± 0.0120	0.0333 ± 0.0020

of the original data. Note that NMF yields a part-based representation, i.e., part of the original images are represented in the dictionary of the decomposition. Therefore, the interpretation of the relevance matrix is related to patterns in the images and not individual pixels. Besides, we evaluated the original USPS data set in order to isolate the effect of the NMF representation and analyse the regularization contribution.

The matrix relevance elements represent patterns and so it is not anymore pixel-wise as it is in the original space. A didactic representation is shown in Fig. 3 in order to help understand this new representation. It shows a schematic of the NMF dictionary and its representation in the relevance matrix. In Fig. 3 (a) two patterns are highlighted from the dictionary and they are related to the two most relevant features shown in the main diagonal in the bar graph Fig. 3 (c) (features 10 and 18). Also, the dictionary of patterns in Fig. 3 (b) shows that there are 8 elements that do not contribute to the representation of the original data, i.e. they are irrelevant as reflected in Fig. 3 (c).

The nuclear norm regularization avoids over-simplifying and preserves the relevant eigenvalues by selecting an appropriate regularization parameter $\beta = -0.05$. It can conserve the relevant information that contributes to get a better classification performance without losing the generalization ability of the GMLVQ model. GMLVQ+nuclear outperforms all other methods and this is statistically significant (see Table 1).

The results show the importance of the preprocessing step done by NMF. It helps reduce the noisy data and it is reflected in the classification performance. Table 1 shows the results obtained using both the original USPS and the projected USPS ($k = 20$) by using NMF. The results show that there is an important contribution of NMF. However, the regularization contribution helps to improve the classification performance in both scenarios.

V. CONCLUSIONS

In this contribution we have proposed a new regularization method for the GMLVQ classifier. It has been shown that the proposed method outperformed the original GMLVQ for all data sets. Also, the proposed method outperformed all the other regularization methods for the USPS data set, and it is tied in the first place with the GMLVQ+log method for the Image Segmentation data set. The nuclear norm works directly over the eigenvalues of the matrix relevance and deletes noisy or small eigenvalues when preventing over-fitting. The results showed that the proposed regularization method prevents over-simplifying over the relevance matrix when $\beta < 0$. However, it is subject to chose a correct regularization parameter such as was previously discussed. A wrong setting of this parameter could result in a lower classification performance. Also, the proposed regularization method could be easily extended to the local matrix model in GMLVQ.

Another issue is that for all experiments presented in this section (except for the toy problem) we have observed that the better value for β is in the range $[-1, 0[$. This confirms the previous finding by other authors that GMLVQ tends to oversimplify its solutions. In addition, it has been shown that GMLVQ can be useful for interpreting patterns from a dictionary of patterns obtained using NMF. In this case the relevance matrix yields a value of relevance for each pattern in the dictionary matrix giving a most comprehensive interpretability of the patterns in the images. The NMF factorization helps delete noisy information as a preprocessing step in image data sets and it contributes to enhance the classification performance of the classifier. In this sense, the proposed regularization contributes to determine which patterns are most useful and allows us to enhance the classification performance.

Acknowledgment

This contribution was funded by CONICYT-CHILE under grant Fondecyt 1171678.

References

[1] T. Kohonen, *Self-organizing maps*, T. Kohonen, Ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997.

[2] D. Nova and P. A. Estévez, “A review of learning vector quantization classifiers,” *Neural Computing and Applications*, vol. 25, no. 3, pp. 511–524, 2013.

[3] M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, and T. Villmann, “Aspects in classification learning-review of recent developments in learning vector quantization,” *Foundations of Computing and Decision Sciences*, vol. 39, no. 2, pp. 79–105, 2014.

[4] A. Bellet, A. Habrard, and M. Sebban, “A survey on metric learning for feature vectors and structured data,” University of Southern California, Tech. Rep., 2013.

[5] P. Schneider, M. Biehl, and B. Hammer, “Adaptive relevance matrices in learning vector quantization,” *Neural Computation*, vol. 21, no. 12, pp. 3532–3561, 2009.

[6] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl, “Limited rank matrix learning, discriminative dimension reduction and visualization,” *Neural Networks*, vol. 26, pp. 159–173, 2012.

[7] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl, “Regularization in matrix relevance learning,” *IEEE Transactions on Neural Networks*, vol. 21, no. 5, pp. 831–840, 2010.

[8] D. Nova and P. A. Estévez, “A study on gmlvq convex and non-convex regularization,” in *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*, E. Merényi, J. M. Mendenhall, and P. O’Driscoll, Eds. Cham: Springer International Publishing, 2016, pp. 305–314.

[9] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *Journal of machine learning research*, vol. 11, no. Aug, pp. 2287–2322, 2010.

[10] ACM-SIGKDD and Netflix, “Soft modelling by latent variables: the nonlinear iterative partial least squares (nipals) approach.” in *In Proceedings of DDD Cup and Workshop*, 2007.

[11] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[12] N. Srebro and A. Shraibman, “Rank, trace-norm and max-norm,” in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 545–560.

[13] K. Mohan and M. Fazel, “Iterative reweighted algorithms for matrix rank minimization,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3441–3473, 2012.

[14] A. S. Lewis, “Derivatives of spectral functions,” *Mathematics of Operations Research*, vol. 21, no. 3, pp. 576–588, 1996.

[15] H. S. Sendov, “The higher-order derivatives of spectral functions,” *Linear algebra and its applications*, vol. 424, no. 1, pp. 240–281, 2007.

[16] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[17] Y.-X. Wang and Y.-J. Zhang, “Nonnegative matrix factorization: A comprehensive review,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2013.